Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **NOV 2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**
4. TITLE AND SUBTITLE **Adapting Autonomous Behavior Based on an Estimate of an Operator's Trust**		5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory ,Navy Center for Applied Research in AI,Washington,DC,20375**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

in AAAI Fall Symposium, Arlington, VA, 13-15 Nov 2014.

14. ABSTRACT

Robots can be added to human teams to provide improved capabilities or to perform tasks that humans are unsuited for. However, in order to get the full benefit of the robots the human teammates must use the robots in the appropriate situations. If the humans do not trust the robots, they may underutilize them or disuse them which could result in a failure to achieve team goals. We present a robot that is able to estimate its trustworthiness and adapt its behavior accordingly. This technique helps the robot remain trustworthy even when changes in context, task or teammates are possible.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**2**	

Adapting Autonomous Behavior Based on an Estimate of an Operator's Trust

Michael W. Floyd and **Michael Drinkwater**
Knexus Research Corporation
Springfield, Virginia, USA
{*michael.floyd, michael.drinkwater*}@*knexusresearch.com*

David W. Aha
Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory (Code 5514)
Washington, DC, USA
david.aha@nrl.navy.mil

Abstract

Robots can be added to human teams to provide improved capabilities or to perform tasks that humans are unsuited for. However, in order to get the full benefit of the robots the human teammates must use the robots in the appropriate situations. If the humans do not trust the robots, they may underutilize them or disuse them which could result in a failure to achieve team goals. We present a robot that is able to estimate its trustworthiness and adapt its behavior accordingly. This technique helps the robot remain trustworthy even when changes in context, task or teammates are possible.

1 Introduction

The addition of robots to human teams can be advantageous if the robots provide sensory capabilities that the humans do not have or can perform tasks the humans are unable to. In a military or search and rescue domain, a robot might have a built-in suite of sensors for detecting hazards in the environment or be able to perform tasks that would be too dangerous for a human to attempt. The success of a human-robot team, and the safety of the team members, will likely be dependant on how well the robots are used to achieve team goals. However, in order for the human teammates to use a robot they must *trust* it.

In a human-robot team, a robot will have certain responsibilities and may be assigned tasks by human teammates. An autonomous or semi-autonomous robot will have direct control over how a task is performed (e.g., the path it uses when moving, how it navigates around obstacles). If the way the robot performs a task is different from how the human teammates would like the task to be performed, the humans may interpret this as poor performance and lose trust in the robot. Lost trust can result in reduced use (e.g., only using the robot for simple tasks), disuse (e.g., the humans perform all of the tasks), or excessive monitoring of the robot's performance.

Ideally, a robot would be designed in such a way that it engenders trust with humans. However, this might be impractical if the way the teammates measure trust is time-dependent, task-dependent, or user-dependent (Desai et al. 2013). For example, one teammate might prefer the robot move between two locations quickly whereas another might prefer the robot move without bumping into any objects. The preferences of the first teammate might change if the robot was transporting delicate cargo whereas the preferences of the second teammate might change in an emergency situation. Without having knowledge about all potential teammates, all possible tasks, and every context under which those tasks will be performed, it would not be possible to program a robot to be trustworthy in every situation.

Our work looks at how a robot can estimate its trustworthiness and adapt its behavior accordingly. Using such an approach, the robot can continuously monitor its trustworthiness and respond to changing contexts, tasks and teammates. In some situations, to behave in a more trustworthy manner the robot might slightly refine its behavior whereas in other situations it might drastically change its behavior. In the remainder of this paper we will describe how the robot can measure a teammate's trust and adapt to it, and also provide directions for this work.

2 Inverse Trust and Behavior Adaptation

Traditional trust metrics measure how much trust an agent should have in another agent and generally use information from past interactions (Sabater and Sierra 2005). Instead, we are looking at an *inverse trust metric* where an agent (the robot) estimates how much trust another agent has in it. One option would be to get direct feedback from the agent about the trustworthiness of the robot (Kaniarasu et al. 2013; Muir 1987). However, this might not be possible in time-critical situations or where there will be a significant delay between opportunities for receiving feedback from teammates.

In our approach, the robot infers a teammate's trust based on the assumption that trust is related to the robot's performance. Although many factors have been found to influ-

ence human-robot trust, the robot's performance has been found to be the strongest indicator (Hancock et al. 2011; Carlson et al. 2014). The robot receives commands from a single teammate, called the *operator*, and performs autonomous behavior to complete the assigned task. If the robot completes the assigned task it assumes the operator's trust is increasing, whereas it assumes the operator's trust is decreasing if it fails to complete the task or is interrupted by the operator.

If the trust decreases past a threshold, called the *untrustworthy threshold*, the robot modifies its behavior in an attempt to perform a more trustworthy behavior. The robot's behavior B has n *modifiable components* and the currently selected values for each of the modifiable components $(c_1, c_2, \ldots, c_n)$ characterize how the robot will behave ($B = \langle c_1, c_2, \ldots, c_n \rangle$). These modifiable components could include the algorithms being used (e.g., switching between two path planning algorithms), the parameter values the robot uses, or the data that is used (e.g., using a different map of the environment). The adaptation, which changes the robot's behavior from B to $B^{'}$, can involve searching through possible behaviors (Floyd, Drinkwater, and Aha 2014a) or using trustworthy behaviors from previous adaptations (Floyd, Drinkwater, and Aha 2014b).

3 Discussion

Our work has focused on an approach for inverse trust estimation and behavior adaptation that does not rely on an explicit model of the operator's preferences and requires minimal interaction with the operator. This reduces the background knowledge required by the robot but is restrictive since it limits the amount of information that can be used in the inverse trust estimation and relies on less efficient behavior search techniques. We plan to allow the robot to use additional information, if it is available, to improve its behavior adaptation. This could include feedback from the operator, dialog between the robot and the operator, or background knowledge about the team objectives.

We have examined situations of undertrust, where the robot should increase the operator's trust, but we have not examined overtrust, where the operator trusts the robot even when it is performing poorly. In these situations the robot should identify when it is behaving poorly and notify the operator. This provides a level of transparency between the robot and the operator, and allows the human teammates to have a better understanding of the capabilities of the robot.

Additionally, we would like to examine having the robot reason about its goals, the goals of the operator, and the goals of the team. By examining how these goals align, the robot can verify it is working towards the correct goals and identify when sudden goal changes occur (e.g., an emergency situation causes the team to abandon their current goal and evacuate the area). Goal reasoning would also allow the robot to build trust models for each goal so that it can quickly switch to goal-specific trustworthy behaviors.

Acknowledgments

Thanks to the Naval Research Laboratory and the Office of Naval Research for supporting this research.

References

Carlson, M. S.; Desai, M.; Drury, J. L.; Kwak, H.; and Yanco, H. A. 2014. Identifying factors that influence trust in automated cars and medical diagnosis systems. In *AAAI Symposium on The Intersection of Robust Intelligence and Trust in Autonomous Systems*, 20–27.

Desai, M.; Kaniarasu, P.; Medvedev, M.; Steinfeld, A.; and Yanco, H. 2013. Impact of robot failures and feedback on real-time trust. In *8th International Conference on Human-Robot Interaction*, 251–258.

Floyd, M. W.; Drinkwater, M.; and Aha, D. W. 2014a. Adapting autonomous behavior using an inverse trust estimation. In *Proceedings of the 14th International Conference on Computational Science and Its Applications*, 728–742.

Floyd, M. W.; Drinkwater, M.; and Aha, D. W. 2014b. How much do you trust me? Learning a case-based model of inverse trust. In *22nd International Conference on Case-Based Reasoning*.

Hancock, P. A.; Billings, D. R.; Schaefer, K. E.; Chen, J. Y.; De Visser, E. J.; and Parasuraman, R. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53(5):517–527.

Kaniarasu, P.; Steinfeld, A.; Desai, M.; and Yanco, H. A. 2013. Robot confidence and trust alignment. In *8th International Conference on Human-Robot Interaction*, 155–156.

Muir, B. M. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27(56):527–539.

Sabater, J., and Sierra, C. 2005. Review on computational trust and reputation models. *Artificial Intelligence Review* 24(1):33–60.